WATER RESOURCES RESEARCH, VOL. 44, W05406, doi:10.1029/2007WR006392, 2008

A relational model for environmental and water resources data

Jeffery S. Horsburgh,[1] David G. Tarboton,[1] David R. Maidment,[2] and Ilya Zaslavsky[3]

Received 30 July 2007; revised 24 January 2008; accepted 13 February 2008; published 8 May 2008.

[1] Environmental observations are fundamental to hydrology and water resources, and the way these data are organized and manipulated either enables or inhibits the analyses that can be performed. The Observations Data Model presented here provides a new and consistent format for the storage and retrieval of point environmental observations in a relational database designed to facilitate integrated analysis of large data sets collected by multiple investigators. Within this data model, observations are stored with sufficient ancillary information (metadata) about the observations to allow them to be unambiguously interpreted and to provide traceable heritage from raw measurements to useable information. The design is based upon a relational database model that exposes each single observation as a record, taking advantage of the capability in relational database systems for querying based upon data values and enabling cross-dimension data retrieval and analysis. This paper presents the design principles and features of the Observations Data Model and illustrates how it can be used to enhance the organization, publication, and analysis of point observations data while retaining a simple relational format. The contribution of the data model to water resources is that it represents a new, systematic way to organize and share data that overcomes many of the syntactic and semantic differences between heterogeneous data sets, thereby facilitating an integrated understanding of water resources based on more extensive and fully specified information.

Citation: Horsburgh, J. S., D. G. Tarboton, D. R. Maidment, and I. Zaslavsky (2008), A relational model for environmental and water resources data, *Water Resour. Res.*, *44*, W05406, doi:10.1029/2007WR006392.

1. Introduction

[2] Environmental observations are fundamental to hydrology and water resources, and the manner in which the data are collected, organized, and manipulated either enables or inhibits their scientific analysis [*Pokorný*, 2006; *Tomasic and Simon*, 1997]. When scientists and engineers want to search for and use environmental observations data, they are generally faced with the following problems [*Tomasic and Simon*, 1997]: (1) data are not sufficient or do not exist; (2) data are not published and are hard to locate; (3) data are not easy to access, they are either private or expensive, or require costly preprocessing before they can be used; (4) data are not easy to use because they are inconsistent or noncompatible; and (5) data are not adequately documented. Addressing these issues is one of the main challenges influencing recent developments in environmental information systems, which include water resources and hydrologic information systems [*Pokorný*, 2006; *Bouganim et al.*, 2001].

[3] Even for data sets that have been published for widespread use, points three through five above still apply. Generally, data sets published on public Web sites are in file-based systems that are different syntactically (e.g., file types, file formats, and data structure) and semantically (e.g., variable names, units, and descriptive metadata) from one data source to the next. In accessing these data archives, users are faced with the daunting task of navigating through directories and supporting files to find all of the metadata necessary for interpreting and using the data. There is a fundamental need within the hydrologic and environmental engineering communities for new, scientific methods to organize and utilize observational data that overcome the syntactic and semantic heterogeneity in data from different experimental sites and sources and that allow data collectors to publish their observations so that they can easily be accessed and interpreted by others. This need is being driven by the ever increasing number of environmental observations being produced as sensor technology improves, as the number, size, and complexity of environmental monitoring programs grow (including efforts to establish a national network of large-scale environmental observatories), and as engineers and scientists realize that it is as important to characterize the environment with observations as it is to describe it with models and simulations. It is critical that the data, when published, be carefully annotated with metadata so that they can be unambiguously interpreted and used.

[4] In this paper we present a logical database design for the Observations Data Model (ODM) that advances the information science knowledge base of water resources research. We describe a relational model that eases access to and manipulation of time series of observations from experimental sites and watersheds and facilitates data pub-

[1]Utah Water Research Laboratory, Utah State University, Logan, Utah, USA.
[2]Center for Research in Water Resources, University of Texas at Austin, Austin, Texas, USA.
[3]San Diego Supercomputer Center, University of California, San Diego, La Jolla, California, USA.

0043-1397/08/2007WR006392$09.00

lishing, querying, retrieval, and analysis among domains and investigators. This design identifies the entities, attributes, and relationships required to represent observations, but it is independent of its physical implementation (i.e., it can be implemented within any relational database management system). This system has been implemented and used to publish a wide range of environmental data at 11 test bed sites that are part of an effort to advance environmental observatory design (http://www.watersnet.org/wtbs/index.html). The experience in implementing this model at these 11 sites has demonstrated the generality and effectiveness of ODM.

[5] ODM is focused on observations made at a point, such as those made at a streamflow gauge or a stationary weather station, although observations recorded from moving platforms or along routes can also be represented by treating location as an observation. The representation of spatially distributed data in ODM is limited to the presentation of time series of point observations that are at different spatial locations. ODM does not include raster data sets, for which we envision a different data model being developed. However, distributed time series data (e.g., time series of raster data sets such as weather radar observational grids) can be represented within ODM by using grid cell centers as observation sites.

[6] ODM is the result of an effort to create a generic model of observational data from a range of water resources disciplines (hydrology, environmental engineering, meteorology, etc.) and to accommodate a range of different variables (precipitation, streamflow, water quality). The model has drawn upon input from community surveys and reviews [*Bandaragoda et al.*, 2005; 2006; D. G. Tarboton, Review of proposed CUAHSI hydrologic information system hydrologic observations data model, Utah State University, 5 May 2005, http://www.engineering.usu.edu/cee/faculty/dtarb/HydroObsDataModelReview.pdf]. ODM has been applied to physical and chemical data from water systems, climate and weather observations, and aquatic biology measurements such as species distributions, and it is this flexibility that is largely responsible for its utility. ODM's ability to store and enable access to similarly formatted data and metadata from multiple domains, for example streamflow data and climate data for inputs to a hydrologic model, can greatly enhance the use of these data and can result in significant time savings and value added to the data. Additionally, the consistent format for data and metadata that ODM provides enables the development of standardized software applications on top of ODM. ODM enables easy and automated access to the data through a relational database management system, which enables multiple software developers to create compatible applications as well as the reuse of code for standard tasks such as data discovery and retrieval.

[7] Additionally, ODM represents a new opportunity for many within the water resources community to approach the management, publication, and analysis of their data systematically, i.e., moving from collections of ASCII text or spreadsheet files to a relational data model that removes the burden of learning and interpreting diverse file formats from the data end user. Systematic data management using relational database systems has advanced data mining, predictive modeling, and deviation detection within the business community, where most operational data is stored in relational databases due to their reliability, scalability, available tools, and performance [*Connolly and Begg*, 2005]. The systematic data analysis capabilities that a relational data model enables have the potential to stimulate similar advances in the water resources area.

[8] In this paper we describe the structure and features of ODM and discuss its implementation for data management in prototype environmental observatories. Section 2 discusses existing standards for environmental observations data. Section 3 describes the requirements considered in designing ODM. Section 4 gives the structure of ODM and describes some of its features. Section 5 provides examples of water resources data that have been incorporated into ODM, and Section 6 discusses the implementation of ODM within a national network of environmental observatory test beds.

2. Existing Standards for Environmental Observations

[9] Much work has already been done to develop standards for exchanging information describing the collection, analysis, and reporting of environmental data. The Environmental Data Standards Council (EDSC) has developed a set of Environmental Sampling, Analysis, and Results Data Standards specifically for this purpose [*Environmental Data Standards Council*, 2006]. A similar standard has been developed by the National Water Quality Monitoring Council (NWQMC) specifically for water quality data elements [*National Water Quality Monitoring Council*, 2006], and the Open Geospatial Consortium (OGC) has developed a best practices document called ''Observations and Measurements'' that describes terminology and presents a framework and encoding for measurements and relationships between them (OGC Best Practices Document OGC 05-087r4, version 0.14.7, available at http://portal.opengeospatial.org/files/?artifact_id=17038). These standards are focused primarily on the data elements required to facilitate the exchange of environmental observations without considering the format for persistent data storage such as in a relational database. In designing ODM, we strove to include the most important attributes of observations from these standards in a logical data model design that can be physically implemented in relational database management systems.

[10] It is important to note that ODM's purpose is to manage the storage and retrieval of observations data as part of a broader hydrologic information system (HIS) that also provides data discovery, analysis, and exchange capability through software applications built on top of ODM. For example, within the HIS being developed by the Consortium of Universities for the Advancement of Hydrologic Sciences, Inc (CUAHSI), the main mechanism for the exchange of environmental observations is the WaterOneFlow Web services (http://www.cuahsi.org/his/webservices.html). Web services are applications that provide the ability to pass information between computers over the Internet [*Goodall et al.*, 2008]. The WaterOneFlow Web services transmit data extracted from an ODM database encoded as eXtensible Markup Language (XML) and formatted using an XML schema called WaterML (Open Geospatial Consortium, Inc., CUAHSI WaterML, OGC Discussion Paper OGC 07-041r1, version 0.3.0, available

Table 1. Observations Data Model Attributes Associated With an Observation

Attribute	Definition
Value	The observation value itself
Accuracy	Quantification of the measurement accuracy associated with the observation value
Date and Time	The date and time of the observation (including time zone offset relative to UTC and daylight savings time factor)
Variable Name	The name of the physical, chemical, or biological quantity that the value represents (e.g. streamflow, precipitation, water quality)
Location	The location at which the observation was made (e.g. latitude and longitude)
Units	The units (e.g. m or m^3/s) and unit type (e.g. length or volume/time) associated with the variable
Interval	The interval over which each observation was collected or implicitly averaged by the measurement method and whether the observations are regularly recorded on that interval
Offset	Distance from a reference point to the location at which the observation was made (e.g., 5 m below water surface)
Offset Type/ Reference Point	The reference point from which the offset to the measurement location was measured (e.g., water surface, stream bank, snow surface)
Data Type	An indication of the kind of quantity being measured (e.g., an instantaneous or cumulative measurement)
Organization	The organization or entity providing the measurement
Censoring	An indication of whether the observation is censored or not
Data Qualifying Comments	Comments accompanying the data that can affect the way the data is used or interpreted (e.g., holding time exceeded, sample contaminated, provisional data subject to change, etc.)
Analysis Procedure	An indication of what method was used to collect the observation (e.g., dissolved oxygen by field probe or dissolved oxygen by Winkler Titration)
Source	Information on the original source of the observation (e.g. from a specific instrument or investigator third-party database)
Sample Medium	The medium in which the sample was collected (e.g., water, air, sediment, etc.)
Quality Control Level	An indication of the level of quality control the data has been subjected to (e.g., raw data, checked data, derived data)
Value Category	An indication of whether the value represents an actual measurement, a calculated value, or is the result of a model simulation

at http://portal.opengeospatial.org/files/?artifact_id=21743). This separation between content (i.e., the data stored in an ODM database) and presentation (i.e., the format of the data when it is transmitted) is an important aspect of the overall HIS design.

3. ODM Design Requirements

[11] An observation is an event that results in a value describing some phenomenon (Open Geospatial Consortium, Inc., Observations and Measurements, OGC Best Practices Document OGC 05–087r4, version 0.14.7, available at http://portal.opengeospatial.org/files/?artifact_id=17038). Observation values are not self describing, and, because of this, interpretation of a particular set of observations requires contextual information, or metadata. Metadata is the descriptive information about data that explains the measurement attributes, their names, units, precision, accuracy, and data layout, as well as the data lineage describing how the data was measured, acquired, or computed [*Gray et al.*, 2005]. The importance of recording fundamental metadata to help others discover and access data products is well recognized [*Bose*, 2002; *Michener et al.*, 1997; *Gray et al.*, 2005]. ODM was designed to store environmental observations along with sufficient metadata to provide traceable heritage from raw measurements to usable information, allowing observations stored in ODM to be unambiguously interpreted and used.

[12] Environmental observations are identified by the following fundamental characteristics: (1) the location at which the observations were made (space), (2) the date and time at which the observations were made (time), and (3) the type of variable that was observed, such as streamflow, water quality concentration, etc. (variable). In addition to these fundamental characteristics, there are many other attributes that provide additional information necessary for interpretation of observational data. These include the methods used to make observations, qualifying comments about the observation, and information about the organization that made the observation.

[13] Table 1 presents general attributes that are important in interpreting and establishing the provenance of an observation. This list of attributes was compiled from comments received from a community review of a preliminary version of ODM (http://www.engineering.usu.edu/cee/faculty/dtarb/HydroObsDataModelReview.pdf). All of the information contained in Table 1, except for the value of the observation itself, can be considered metadata. The ODM logical data model given in the following section has been designed to store observation values and their supporting metadata in a structured way.

4. ODM Logical Data Model

[14] The logical data model for ODM is shown in Figure 1. The DataValues table at the center stores the numeric values for observations and links (foreign keys) to all of the data value level attributes. Most of the attribute details are stored in the tables surrounding the DataValues table to avoid redundancy. The relationships between tables are shown, along with all of the required primary and foreign keys. Each of these relationships has a name, which is indicated by a text label, and a directionality that is indicated by an arrow. For example, the relationship between the Sources table and the DataValues table is named "Generate" and has directionality that points from the Sources table to the DataValues table. This indicates that data sources generate data values. Additionally, the cardinality, or numeric relationship between entities in each of the tables, is shown at either end of each of the relationship lines. For example, the relationship line between the Variables and DataValues tables has "1..1" at the Variables end, and "0..*" at the DataValues end, indicating that there is one and only one variable associated with 0 or many

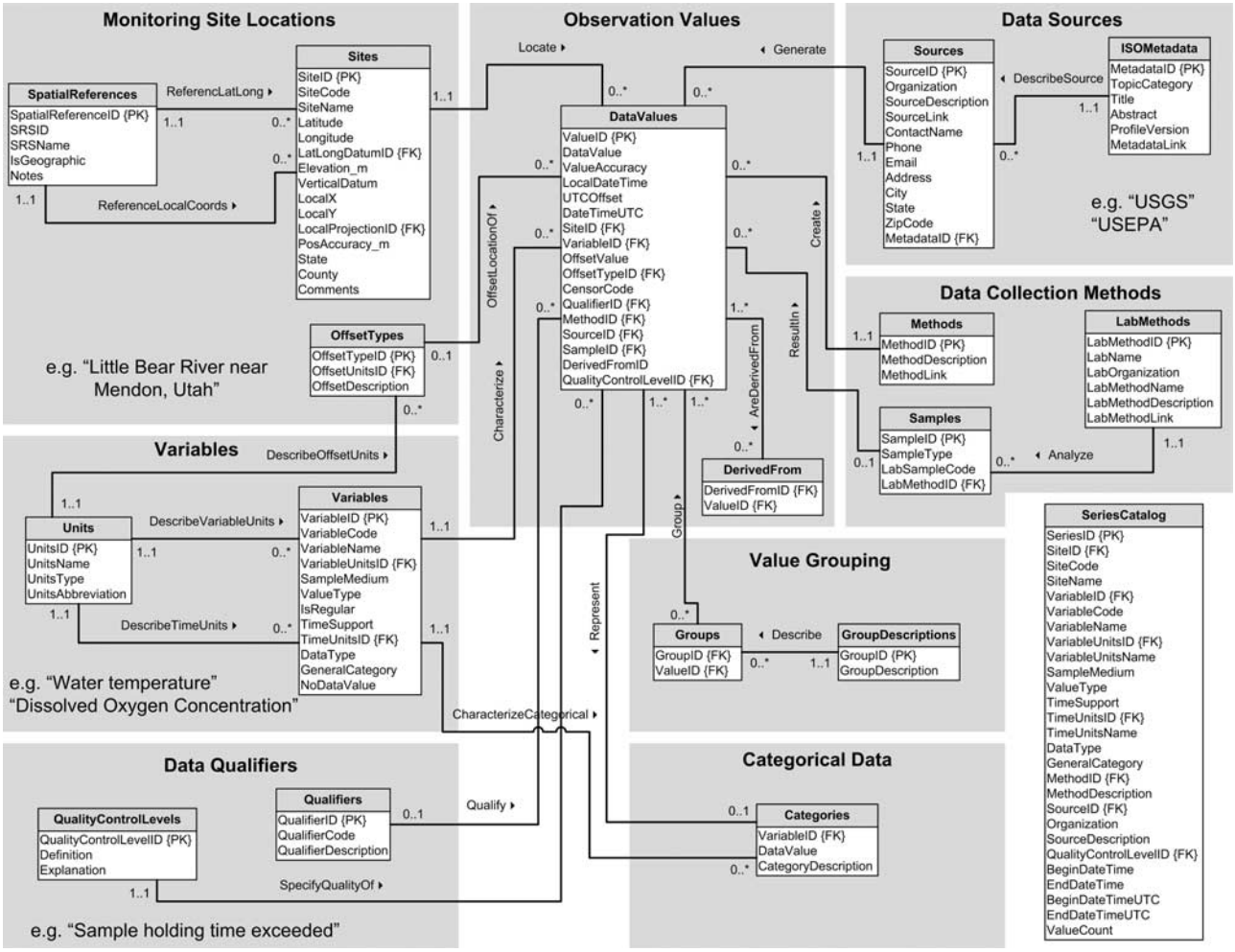


Figure 1. Observations Data Model (ODM) logical data model. The primary key field for each table is designated with a {PK} label. Foreign keys are designated with a {FK} label. The lines between tables show relationships with cardinality indicated by numbers and labeled with the name and directionality of the relationship.

DataValues (i.e., there is a one-to-many relationship between variables and data values) and that variables characterize data values. The subsections that follow describe how ODM encodes observations and their supporting metadata. Readers are referred to D. G. Tarboton et al. (CUAHSI community Observations Data Model (ODM) design specifications document: Version 1.0, http://water.usu.edu/cuahsi/odm/files/ODM1.pdf) for the complete ODM design specifications and data dictionary.

4.1. Monitoring Site Geography, Location, and Offset

[15] Within ODM, the geographic location of monitoring sites is specified through latitude and longitude coordinates as well as elevation information recorded in the Sites table. Additionally, ODM provides the option to specify local coordinates, which may be in a standard geographic projection (e.g., universal transverse Mercator) or a locally defined coordinate system specific to a study area. Both the spatial reference system associated with the horizontal and vertical coordinates and the accuracy with which the location of a monitoring site is known can be quantified within ODM. The field PosAccuracy_m is a numeric value intended to specify the uncertainty in the spatial location information.

[16] Each monitoring site has a unique identifier that can be logically linked to one or more objects in a Geographic Information System (GIS) data model. Figure 2 depicts relationships between monitoring sites within an ODM database and points in a GIS data model. The GIS data model depicted in Figure 2 is Arc Hydro, which is a data structure for linking stream networks, monitoring points and watersheds within a GIS [*Maidment*, 2002]. This linkage between unique monitoring site identifiers and GIS object identifiers is generic and suitable for use with any geographic data model that includes the location of monitoring sites. For example, a linear referencing system on a river network, such as the National Hydrography Dataset (see NHDPlus user guide, ftp://ftp.horizon-systems.com/NHDPlus/documentation/NHDPLUS_UserGuide.pdf), might be used to specify the location of a site on a river network. Information from direct addressing relative to hydrologic objects, such as position of a stream gauge along a stream reach, is often of greater value to a user than latitude and longitude information [*Maidment*, 2002].

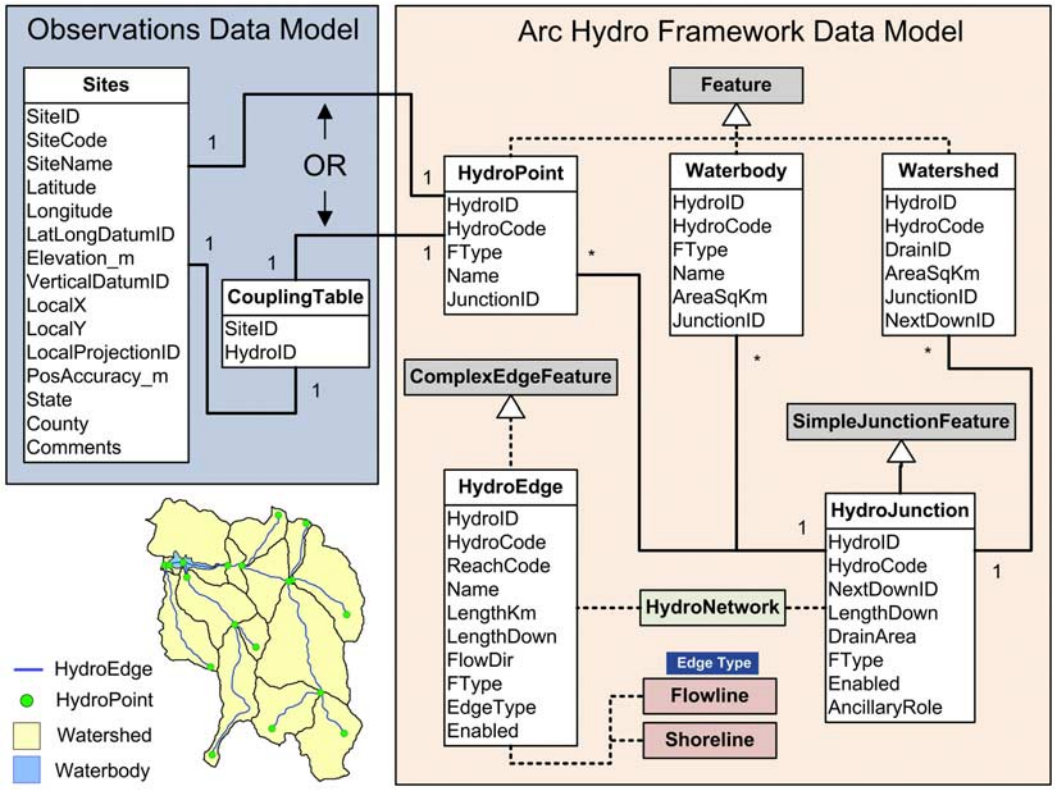


Figure 2. Arc Hydro Framework Data Model and Observations Data Model related through the SiteID field in the Sites table.

[17] The location at which observations were made may also be qualified by an offset, which is used to record the location of an observation relative to an appropriate local reference point, such as depth below the water surface. In some cases, such local reference is required for proper interpretation of the data. For example, observations of water temperature or dissolved oxygen may be made at a number of different depths at a location within a water body. The offset would be used to quantify the depth of each measurement below the surface. Within ODM, an offset is specified by a numeric value that is the offset distance, the units of the offset, and an offset description that defines the type of offset (e.g., below the water surface or above ground level).

4.2. Variable Information

[18] The variables that can be represented in ODM range from hydrologic variables such as discharge and gauge height to water quality variables such as nutrient and sediment concentrations to meteorological variables such as air temperature and precipitation as well as many others. The most fundamental attribute of an environmental variable is its name (e.g., discharge or temperature), but there are several other variable attributes recorded in ODM that are important, including: (1) the units of the observations for a variable (e.g., $m^3\ s^{-1}$), (2) the medium in which the observations are made (e.g., surface water or sediment), (3) the regularity with which observations are made, (4) the support, spacing, and extent of observations, and (5) the nature of the observation as an actual measurement (e.g., stage) or a derived value (e.g., discharge derived from stage). All of this information is represented at the variable level within ODM.

4.2.1. Time Support, Spacing, and Extent

[19] To interpret values that comprise a time series or set of observations, it is important to know the timescale information associated with the values. *Blöschl and Sivapalan* [1995] review the important issues. Any set of observations is quantified by a scale triplet comprising support, spacing, and extent. Extent is the full range of time over which the observations occur, spacing is the time between observations, and support is the averaging interval implicit in any observation. In ODM, the time support associated with observations is specified by a numeric value that quantifies the support and an indication of the units associated with the support value. Extent and spacing are properties of multiple observations and are defined by the set of dates and times associated with the observations. Dates and times associated with observations are stored in local time (in the time zone in which the observation was made), UTC time, and ODM also stores the UTC offset to ensure that dates and times are unambiguous.

4.2.2. Data Types

[20] The environmental processes that we wish to characterize through observation may be dynamic and continuous in nature, but our ability to measure them is constrained to particular instants or intervals of time. To interpret environmental observations, it is important to know whether an observation is an instantaneous result, such as in the case of water quality variables where a sample is collected at an instant in time, or whether the observation is a cumulative or incremental value resulting from a measurement device

Table 2. Data Types That Can Be Represented Within Observations Data Model

Data Type	Description	Example
Continuous	The phenomenon, such as streamflow, Q(t) is specified at a particular instant in time and measured with sufficient frequency (small spacing) to be interpreted as a continuous record of the phenomenon.	Fifteen minute observations of discharge at a stream gauge station.
Sporadic	The phenomenon is sampled at a particular instant in time but with a frequency that is too coarse for interpreting the record as continuous. This would be the case when the spacing is significantly larger than the support and the timescale of fluctuation of the phenomenon.	Infrequent water quality samples that characterize nutrient concentrations.
Cumulative	The data represents the cumulative value of a variable measured or calculated up to a given instant of time: $V(t) = \int_0^t Q(\tau)d\tau$, where τ represents time in the integration over the interval [0, t].	Cumulative volume of flow or cumulative precipitation.
Incremental	The data value represents the incremental value of a variable over a time interval Δt: $\Delta V(t) = \int_t^{t+\Delta t} Q(\tau)d\tau$.	Incremental volume of flow or incremental precipitation.
Average	The data value represents the average over a time interval, such as daily mean discharge or daily mean temperature: $\bar{Q}(t) = \frac{\Delta V(t)}{\Delta t}$. The averaging interval is quantified by time support in the case of regular data and by the time interval from the previous data value at the same position for irregular data.	Daily mean discharge or daily mean air temperature.
Maximum	The data value is the maximum value occurring at some time during a time interval. ODM adopts the convention that the time interval is the time support for regular data and the time interval from the previous data value at the same position for irregular data.	Annual maximum discharge or daily maximum air temperature.
Minimum	The data value is the minimum value occurring at some time during a time interval. The time interval is defined similarly to Maximum data.	The 7-day low flow for a year or daily minimum air temperature.
Constant Over Interval	The data value is a quantity that can be interpreted as constant over the time interval from the previous measurement.	Discharge from a control structure that does not change unless a gate is moved or reset.
Categorical	The value stored is a numerical value that represents a categorical rather than continuous valued quantity. Each category is represented by a numeric value, and the mapping from numeric values to categories is stored in ODM.	Weather observations such as "Cloudy" or "Partly Cloudy."

such as a rain gauge that accumulates a quantity over time. In ODM this information is referred to as the data type and is recorded in the DataType attribute in the Variables table. Table 2 lists the major data types that can be represented within ODM. This list expands upon the data types listed by *Maidment* [2002], and it is anticipated that as more data types are incorporated into specific ODM instances that this list will grow.

4.2.3. Samples and Methods

[21] The method used to make a measurement is important for its interpretation. Within ODM, individual observation values can be associated with a record in the Methods table that describes how a physical observation was made or collected. Descriptive information about each measurement method can be stored and can include specific and detailed information about the technique or equipment used. In the case of observations derived from laboratory samples, ODM provides the additional feature of storing information in the Samples table to link individual observations to the specific physical samples analyzed in a laboratory. Details about the laboratory methods and protocols used in analyzing the samples can be stored in the LabMethods table.

4.3. Quality Control

[22] Data versioning and quality control are key concepts in environmental data management where raw data streams in from in situ sensors through telemetry networks. Raw sensor data can contain a variety of errors caused by equipment malfunction, instrument drift, improper calibration, vandalism, or other causes. In most cases, raw sensor data are not useful for defensible scientific analyses until they have been filtered through a quality control process. To accommodate quality control measures and data versioning, each observation stored in ODM is assigned a quality control level that indicates the level of quality control to which a value has been subjected. The quality control levels used within ODM are stored in the QualityControlLevels table and have been adapted from those used by other earth observatory projects and communities [*Ahern*, 2004; NASA, Committee on Data Management, Archiving, and Computing (CODMAC) data level definitions, http://science.hq.nasa.gov/research/earth_science_formats.html] so that ODM is consistent with these other efforts. The definitions for the quality control levels used by ODM are listed in Table 3.

4.4. Value Accuracy

[23] Each observation stored in ODM can be attributed with an indication of the accuracy of the observation. This attribute is a numeric value that quantifies the total measurement accuracy defined as the nearness of a measurement to the true or standard value. The value accuracy quantifies the uncertainty of the measurement due to errors in both bias and precision. In practice, since the true value is not known, the value accuracy should be estimated based on knowledge of the instrument accuracy, measurement method, and operational environment. In some cases, it is possible to quantify precision by statistical analysis of the scatter associated with repeated measurements and to quantify bias through comparison to specially designed unbiased measurements. Value accuracy can then be estimated by

Table 3. Quality Control Levels in Observations Data Model

Level	Description	Example
0	Raw and unprocessed data and data products that have not undergone quality control. Depending on the variable, data type, and data transmission system, raw data may be available within seconds or minutes after the measurements have been made.	Real-time precipitation, streamflow, and water quality measurements.
1	Quality-controlled data that have passed quality assurance procedures such as routine estimation of timing and sensor calibration or visual inspection and removal of obvious errors.	USGS published daily average discharge records following parsing through USGS quality control procedures.
2	Derived products that require scientific and technical interpretation and may include multiple-sensor data.	Basin average precipitation derived from rain gauges using an interpolation procedure.
3	Interpreted products that require researcher driven analysis and interpretation, model-based interpretation using other data and/or strong prior assumptions.	Basin average precipitation derived from the combination of rain gauges and radar return data.
4	Knowledge products that require researcher driven scientific interpretation and multidisciplinary data integration and include model-based interpretation using other data and/or strong prior assumptions.	Percentages of old or new water in a hydrograph inferred from an isotope analysis.

combining these using a root mean square sum. In other cases, value accuracy will be a more subjective estimate.

[24] Value accuracy is an observation level attribute because it can change with each measurement, dependent on the instrument or measurement protocol. For example, if streamflow is estimated using a V notch weir, it is actually the stage that is measured, with accuracy limited by the precision and bias of the depth recording instrument. The conversion to discharge through the stage-discharge relationship results in greater absolute error for larger discharges. Inclusion of the value accuracy attribute, which will be unknown for many historic data sets because historically accuracy has not been recorded, adds to the size of data in ODM, but provides a way for factoring the accuracy associated with measurements into data analysis and interpretation, a practice that should be encouraged.

4.5. Groups and Derived From Associations

[25] ODM provides the capability to associate observations into logical groups using the Groups and GroupDescriptions tables. Observation groups maintain association between related data values (e.g., all of the temperature observations from a single lake depth profile). Each observation group is identified by a group name and a list of all of the unique ValueIDs for the data values that make up the group. There is no limit to how many observation groups a data value may be associated with.

[26] ODM also provides the capability to store derived quantities (e.g., discharge) and the observations (e.g., stage) from which they were derived. Raw observation values and values derived from raw observations are stored together in the central DataValues table, while the connection between each derived data value and its more primitive raw measurement is preserved in the DerivedFrom table. Derived values may be created by transforming data, for example transforming stage to discharge, or by simply creating a quality-controlled data series from a raw data series. Derived values may be associated with one or many more primitive data values via the DerivedFrom table to, for example, identify the single gauge height value used to estimate an instantaneous discharge value, or the 96 instantaneous discharge values at 15-min intervals that go into an estimate of mean daily discharge. Preserving the relationships between data values and the values from which they were derived is important in maintaining the provenance of observations.

4.6. Qualifying Comments and Censored Data

[27] Many observations are accompanied by comments that qualify how the data should be interpreted or used. These comments are important in stipulating the quality of the data or in flagging potential problems. For example, when sample holding times associated with a particular chemical analysis method are exceeded before a sample is analyzed, the resulting data may be suspect. Data qualifying comments are typically added to such observations by the laboratory that performs the analysis, and it is critical that these comments follow the data wherever they are used. To this end, each individual observation stored within ODM can be qualified by a text comment that describes limitations of, or information about, that observation that are required in interpreting its value and in evaluating its appropriateness for use.

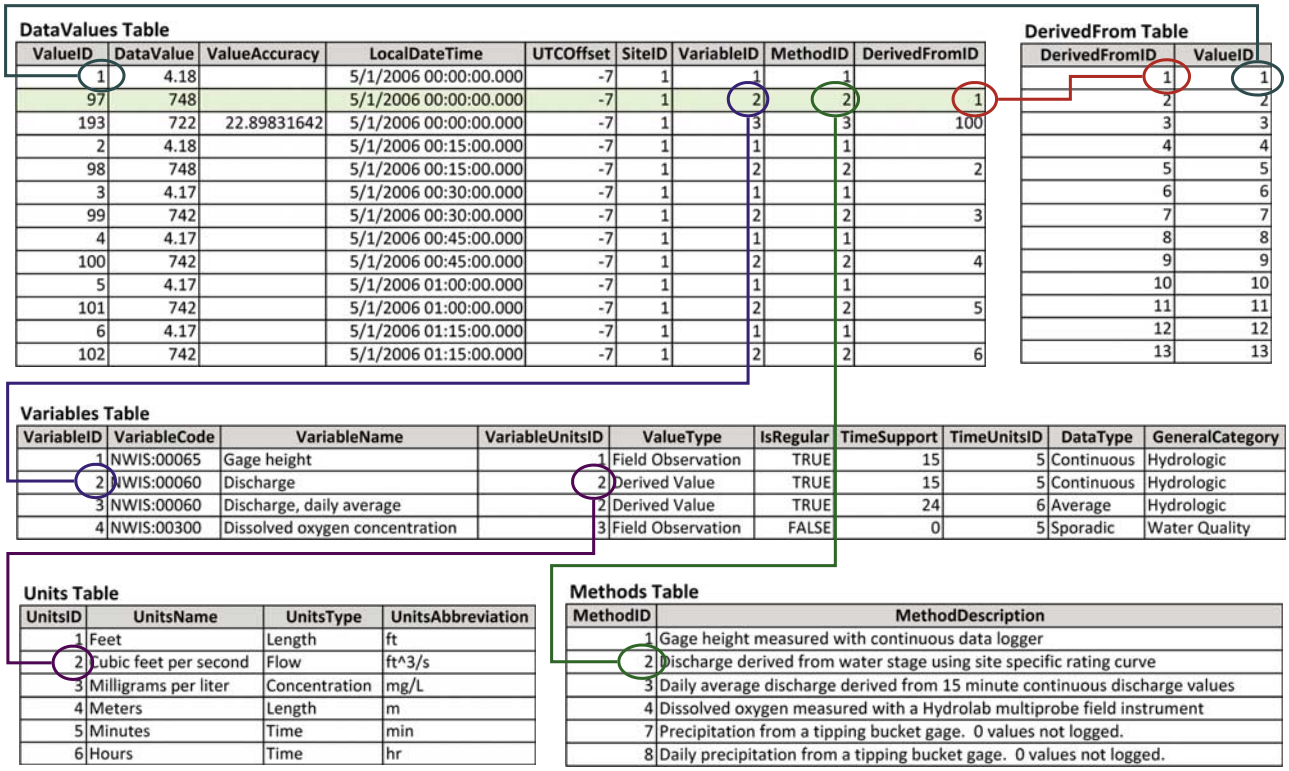
[28] Censored data, or data that are above or below a detection or quantitation limit, are another issue that must be dealt with in storing environmental observations. Within ODM, each individual observation can be qualified by a censor code that indicates whether the true value is greater than or less than the value that is reported. All other values are assumed to be not censored. ODM uses a convention similar to that used by the USGS of recording the censoring level (e.g., the detection limit or the quantitation limit) as the value, preserving this information for data analysis methods that require that the censoring level be known [e.g., *Helsel*, 1990].

4.7. Data Sources

[29] Information about the organization responsible for collecting and analyzing the data is an important part of data provenance. ODM provides a link for each observation in the database to the Sources table that holds information about the organization that originally collected the data.

4.8. Controlled Vocabularies

[30] A controlled vocabulary is a carefully selected list of words and phrases that is used to describe units of information or data. Each of the terms within a controlled vocabulary has a unique and unambiguous definition.



DataValues Table

ValueID	DataValue	ValueAccuracy	LocalDateTime	UTCOffset	SiteID	VariableID	MethodID	DerivedFromID
1	4.18		5/1/2006 00:00:00.000	-7	1	1	1	
97	748		5/1/2006 00:00:00.000	-7	1	2	2	1
193	722	22.89831642	5/1/2006 00:00:00.000	-7	1	3	3	100
2	4.18		5/1/2006 00:15:00.000	-7	1	1	1	
98	748		5/1/2006 00:15:00.000	-7	1	2	2	2
3	4.17		5/1/2006 00:30:00.000	-7	1	1	1	
99	742		5/1/2006 00:30:00.000	-7	1	2	2	3
4	4.17		5/1/2006 00:45:00.000	-7	1	1	1	
100	742		5/1/2006 00:45:00.000	-7	1	2	2	4
5	4.17		5/1/2006 01:00:00.000	-7	1	1	1	
101	742		5/1/2006 01:00:00.000	-7	1	2	2	5
6	4.17		5/1/2006 01:15:00.000	-7	1	1	1	
102	742		5/1/2006 01:15:00.000	-7	1	2	2	6

DerivedFrom Table

DerivedFromID	ValueID
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
11	11
12	12
13	13

Variables Table

VariableID	VariableCode	VariableName	VariableUnitsID	ValueType	IsRegular	TimeSupport	TimeUnitsID	DataType	GeneralCategory
1	NWIS:00065	Gage height	1	Field Observation	TRUE	15	5	Continuous	Hydrologic
2	NWIS:00060	Discharge	2	Derived Value	TRUE	15	5	Continuous	Hydrologic
3	NWIS:00060	Discharge, daily average	2	Derived Value	TRUE	24	6	Average	Hydrologic
4	NWIS:00300	Dissolved oxygen concentration	3	Field Observation	FALSE	0	5	Sporadic	Water Quality

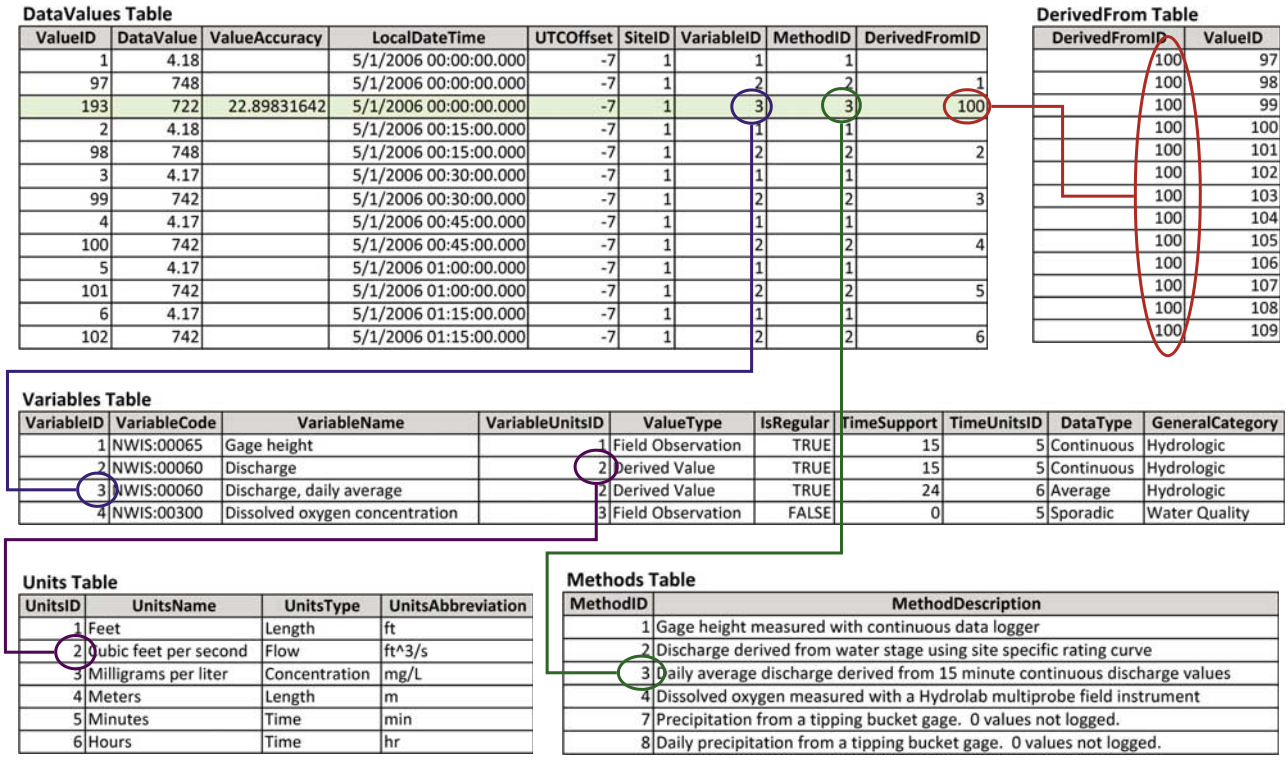
Units Table

UnitsID	UnitsName	UnitsType	UnitsAbbreviation
1	Feet	Length	ft
2	Cubic feet per second	Flow	ft^3/s
3	Milligrams per liter	Concentration	mg/L
4	Meters	Length	m
5	Minutes	Time	min
6	Hours	Time	hr

Methods Table

MethodID	MethodDescription
1	Gage height measured with continuous data logger
2	Discharge derived from water stage using site specific rating curve
3	Daily average discharge derived from 15 minute continuous discharge values
4	Dissolved oxygen measured with a Hydrolab multiprobe field instrument
7	Precipitation from a tipping bucket gage. 0 values not logged.
8	Daily precipitation from a tipping bucket gage. 0 values not logged.

Figure 3. Excerpts from tables illustrating the population of ODM with streamflow gauge height (stage) and discharge data.



DataValues Table

ValueID	DataValue	ValueAccuracy	LocalDateTime	UTCOffset	SiteID	VariableID	MethodID	DerivedFromID
1	4.18		5/1/2006 00:00:00.000	-7	1	1	1	
97	748		5/1/2006 00:00:00.000	-7	1	2	2	1
193	722	22.89831642	5/1/2006 00:00:00.000	-7	1	3	3	100
2	4.18		5/1/2006 00:15:00.000	-7	1	1	1	
98	748		5/1/2006 00:15:00.000	-7	1	2	2	2
3	4.17		5/1/2006 00:30:00.000	-7	1	1	1	
99	742		5/1/2006 00:30:00.000	-7	1	2	2	3
4	4.17		5/1/2006 00:45:00.000	-7	1	1	1	
100	742		5/1/2006 00:45:00.000	-7	1	2	2	4
5	4.17		5/1/2006 01:00:00.000	-7	1	1	1	
101	742		5/1/2006 01:00:00.000	-7	1	2	2	5
6	4.17		5/1/2006 01:15:00.000	-7	1	1	1	
102	742		5/1/2006 01:15:00.000	-7	1	2	2	6

DerivedFrom Table

DerivedFromID	ValueID
100	97
100	98
100	99
100	100
100	101
100	102
100	103
100	104
100	105
100	106
100	107
100	108
100	109

Variables Table

VariableID	VariableCode	VariableName	VariableUnitsID	ValueType	IsRegular	TimeSupport	TimeUnitsID	DataType	GeneralCategory
1	NWIS:00065	Gage height	1	Field Observation	TRUE	15	5	Continuous	Hydrologic
2	NWIS:00060	Discharge	2	Derived Value	TRUE	15	5	Continuous	Hydrologic
3	NWIS:00060	Discharge, daily average	2	Derived Value	TRUE	24	6	Average	Hydrologic
4	NWIS:00300	Dissolved oxygen concentration	3	Field Observation	FALSE	0	5	Sporadic	Water Quality

Units Table

UnitsID	UnitsName	UnitsType	UnitsAbbreviation
1	Feet	Length	ft
2	Cubic feet per second	Flow	ft^3/s
3	Milligrams per liter	Concentration	mg/L
4	Meters	Length	m
5	Minutes	Time	min
6	Hours	Time	hr

Methods Table

MethodID	MethodDescription
1	Gage height measured with continuous data logger
2	Discharge derived from water stage using site specific rating curve
3	Daily average discharge derived from 15 minute continuous discharge values
4	Dissolved oxygen measured with a Hydrolab multiprobe field instrument
7	Precipitation from a tipping bucket gage. 0 values not logged.
8	Daily precipitation from a tipping bucket gage. 0 values not logged.

Figure 4. Excerpts from tables illustrating the population of ODM with daily average discharge derived from 15 min discharge values.

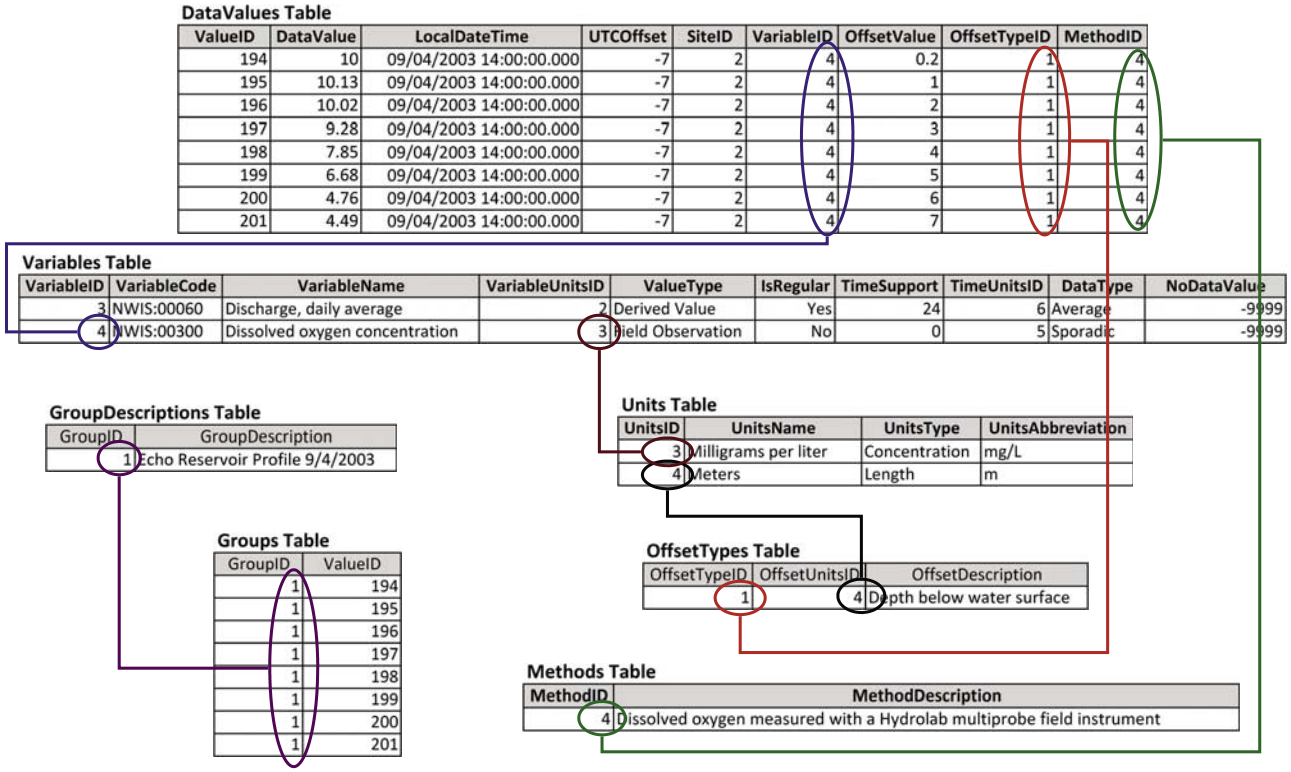


Figure 5. Excerpts from tables illustrating the population of ODM with water chemistry data from a profile in a lake.

ODM imposes controlled vocabularies on some fields within the data model for several reasons. First, the use of controlled vocabularies for elements such as variable and unit names eliminates the use of different terms for the same concept (e.g., "water temperature" versus "temperature, water") and resolves any associated ambiguity. Second, controlled vocabularies can improve the accuracy and performance of searches over fields that could otherwise contain repetitive or ambiguous terms. Additionally, controlled vocabularies form the basis of the metadata within ODM and provide specific language to describe characteristics of the data to aid in its identification, discovery, assessment, and management.

4.9. Data Series

[31] In order to support common data discovery queries that identify which variables have been measured at which locations and for what time periods, we use the concept of a "data series" as an organizing principle within ODM. A data series is a set of observation values of a particular type (e.g., continuously measured water temperature or irregular, instantaneous observations of nitrate concentrations), measured at a single site by a single source using a single method. The ODM SeriesCatalog table maintains a list of all of the data series within the database and essentially performs for an ODM database what a card catalog does for a library. It enables users to search for the data they are looking for as well as providing them with enough information to retrieve the data from the database. This table was designed to satisfy many common data discovery queries such as "which variables have been collected at a particular site" or "which sites have data for a particular variable." Evaluation of these common queries against the SeriesCatalog table rather than against the DataValues table, which holds all of the observation values, significantly simplifies and improves the performance of these queries and facilitates more efficient data discovery.

5. ODM Examples

[32] The examples in the following sections demonstrate the capability of the ODM data model to store different types of point observations. The examples present selected fields and tables chosen to illustrate key capabilities of the data model. These examples are presented using table names and field names shown in Figure 1. For a more in depth listing of ODM examples and a data dictionary that describes in detail all of the tables and fields within ODM, readers are referred to the ODM Design Specifications document (http://water.usu.edu/cuahsi/odm/files/ODM1.pdf). Additional resources, sample databases, and software applications for using ODM can be found on the ODM Web site (http://water.usu.edu/cuahsi/odm/).

5.1. Streamflow: Gauge Height and Discharge

[33] Figure 3 illustrates how both stream gauge height measurements and the associated discharge estimates derived from the gauge height measurements can be stored in ODM. Note that gauge height in feet and discharge in cubic feet per second are both in the same data table but with different VariableIDs that reference the Variables table, which specifies the variable name, units, and other quantities associated with these data values. The link between VariableID in the DataValues table and Variables table is shown. In this example, discharge measurements are de-

rived from gauge height (stage) measurements through a rating curve. The MethodID associated with each discharge record references into the Methods table that describes this and provides a URL that contains metadata details for this method. The DerivedFromID in the DataValues table references into the DerivedFrom table that references back to the corresponding gauge height in the DataValues table from which the discharge was derived.

5.2. Streamflow: Daily Average Discharge

[34] Figure 4 shows excerpts from tables illustrating the population of ODM with both continuous discharge values and derived daily averages. Daily average streamflow is reported as an average of continuous 15 min interval data values. The record giving the single daily average discharge with a value of 722 $ft^3\ s^{-1}$ in the DataValues table has a DerivedFromID of 100. This refers to multiple records in the DerivedFrom table, with associated ValueIDs 97, 98, 99, ... 113 shown. These refer to the specific 15 min discharge values in the DataValues table used to derive the average daily discharge. VariableID in the DataValues table identifies the appropriate record in the Variables table specifying that this is a daily average discharge with units of $ft^3\ s^{-1}$ from UnitsID referencing in to the Units table. MethodID in the DataValues table identifies the appropriate record in the Methods table specifying that the method used to obtain this data value was daily averaging.

5.3. Water Chemistry From a Profile in a Lake

[35] Reservoir profile measurements provide an example of the logical grouping of data values and data values that have an offset in relationship to the location of the monitoring site. These measurements may be made simultaneously (by multiple instruments in the water column) or over a short time period (one instrument that is lowered from top to bottom). Figure 5 shows an example of how these data would be stored in ODM. The OffsetTypes table and OffsetValue attribute are used to quantify the depth offset associated with each measurement. Each of the data values shown has an OffsetTypeID that references into the OffsetTypes table. The OffsetTypes table indicates that for this OffsetType the offset is "Depth below water surface." The OffsetTypes table references into the Units table indicating that the OffsetUnits are meters, so OffsetValue in the DataValues table is in units of meters depth below the water surface.

[36] Each of the data values shown has a VariableID that in the Variables table indicates that the variable measured was dissolved oxygen concentration in units of mg L^{-1}. Each of the data values shown also has a MethodID that in the Methods table indicates that dissolved oxygen was measured with a Hydrolab multiprobe. The combination of the variable name, units, and method are sufficiently general to describe what has been measured. Within the ODM controlled vocabularies, the convention is that the units remain generic, whereas the variable names are more specific. For example, "dissolved phosphorus as P" is a different variable name than "dissolved phosphorus as PO_4," but the units of both are mg L^{-1}.

[37] Additionally, the data values shown are part of a logical group of data values representing the water chemistry profile in a lake. This is represented using the Groups table and GroupDescriptions table. The Groups table associates GroupID 1 with each of the ValueIDs of the data values belonging to the group. A description of this group is given in the GroupDescriptions table.

6. ODM Implementation

[38] As part of the process of planning for a national network of environmental observatories, 11 test bed projects across the United States are focused on developing techniques and technologies for environmental observatories ranging from innovative application of environmental sensors to publishing observations data in common formats that can be accessed by investigators nationwide. The test bed sites are located in a range of environmental conditions from the high Sierra Nevada of California to urban Baltimore, Maryland. Investigators at each of the test beds are participating in the development and deployment of common hydrologic information system capability for publishing observations from each of the test beds. Because a common cyberinfrastructure is being adopted, it is enabling cross-domain analysis within individual test beds as well as cross–test bed sharing and analysis of data. More information about the test beds and the data being collected at each can be found at the following URL (http://www.watersnet.org/wtbs/index.html). The following sections describe how ODM is being used as the basis for the common cyberinfrastructure across the test bed sites and how the issues of heterogeneity in data syntax and semantics are being overcome.

6.1. Overcoming Syntactic Heterogeneity

[39] Within each of the test beds, one barrier in publishing and making use of observational data has been heterogeneity in the syntax of the data. It has been observed, for example, that data downloaded from automated data loggers are formatted differently than data generated as a result of chemical analysis of water samples in a laboratory, and within the test beds, these are only two of a variety of data sources. In addition to these methodological inconsistencies, syntactic heterogeneity within the test beds has also been caused by a proliferation of different file types (e.g., ASCII text files versus Microsoft Excel files), different file formats (e.g., cross-tab tables versus serial lists), as well as other differences that are, in general, a result of investigator preference. Individuals working at the test bed sites all have their own favorite software and file formats in which they choose to work.

[40] ODM has overcome this syntactic heterogeneity by providing a common and encompassing database within which all of the observations, regardless of source, collection method, or original file type and format, can be stored along with their metadata. A variety of software tools have been developed for assisting with and automating the process of loading data into an ODM database. Once data have been loaded from their original format into an ODM database, they are syntactically similar and become available to analytical tools that exploit this format. For example, the WaterOneFlow Web services are the main mechanism for publishing and exchanging observations between test beds. The WaterOneFlow Web services, which have been built to extract data from an ODM database based on a user defined query and transmit it over the Internet, preserve the syntactic homogeneity achieved by loading data into ODM

because the data are transmitted in a single format that is consistent across test beds.

6.2. Overcoming Semantic Heterogeneity

[41] Semantic heterogeneity has been another barrier in the effective publishing and use of observational data that has been addressed within and across the test beds. Semantic heterogeneity refers to the variety in language used to describe observations. Within the test beds, ODM has overcome two different types of semantic heterogeneity: (1) the language used to describe the names of observation attributes and (2) the language used to encode observation attribute values. The first type is general, and is addressed through the standard table and field schema of ODM. For example, within ODM a monitoring location is called a "Site" and all Site attributes are stored in a table called "Sites." In each ODM database, the table names and field/attribute names are consistent and so when investigator data are loaded into ODM they adopt a consistent language.

[42] The second type of semantic heterogeneity is in the attribute values themselves. For example, within ODM, each variable has an attribute called "VariableName" that describes the variable that has been measured. Within the test beds, different investigators use different names for the same constituent (e.g., "water temperature" versus "temperature, water"). These differences are reconciled within ODM through the use of controlled vocabularies. Since the controlled vocabularies within ODM list the terms that are acceptable for use within many fields in the database, only one of the terms describing water temperature would be available in the ODM variable name controlled vocabulary and so when multiple data sets are added to an ODM database they are reconciled through the use of appropriate and consistent controlled vocabulary terms to describe the data. The ODM controlled vocabularies are dynamic and growing in that users can add new terms or edit existing terms by using the functionality on the ODM Web site (http://water.usu.edu/cuahsi/odm/).

6.3. A National Network of Consistent Data

[43] By providing a new method for overcoming the syntactic and semantic heterogeneity in data being collected and published at each of the test bed sites, ODM, along with the WaterOneFlow Web services, has enabled a group of independent test bed investigators working on very different science problems to create a national network of published observational data that enables cross-domain and cross–test bed access to data. The advantages are clear: (1) consistent and fully specified data lead to higher-quality analyses with less uncertainty; (2) the test bed network enabled by ODM is a new data resource for the scientific community; and (3) a standard method for publishing observational data means that the network can grow as more investigators publish their data.

7. Discussion and Conclusions

[44] A data model for storing and managing environmental observations has been presented. The importance of metadata in describing environmental observations data cannot be overstated. It is critical that the data be carefully documented and annotated with metadata so that it can be unambiguously interpreted and used by investigators other than those that collected the data. The colocation of observational data and their associated metadata within a single, integrated ODM database enables easy and automated access.

[45] The reliance of ODM on relational database technology provides several advantages. First, implementation of ODM within a relational database management system enables users to take advantage of the mature technology and advanced tools available in relational database systems. These include data import and export tools, a standardized, high-level query language, and, more recently, tools for advanced data analysis and manipulation such as online analytical processing (OLAP), data mining, and data warehousing.

[46] Next, ODM provides a framework in which data of different types and from disparate sources can be integrated. For example, data from multiple scientific disciplines can be assembled within a single ODM instance (e.g., hydrologic variables, water quality variables, climate variables, etc.). This has been the case at each site within a national network of environmental observatory test beds where publishing observational data using ODM and the WaterOneFlow Web services has enabled both multidisciplinary and cross–test bed access to a national network of consistent data.

[47] The number of characteristics used to describe observations can potentially be large and different across data sources. One significant advantage of ODM is that, along with the observation values, it provides a place to store a standard set of the most commonly used attributes of environmental observations. As with any other model, this representation has some limitations. However, once assembled within ODM, observations can be presented in a consistent way, negating the need for users to learn the diverse data formats of multiple scientific communities. This can be useful when data from multiple disciplines need to be combined into a single analysis or simulation model.

[48] Last, a consistent data model enables the standardization of software application development. These software tools include the WaterOneFlow Web services, data loading and editing tools, and data visualization and retrieval tools. Readers are referred to the CUAHSI HIS Web site for details of these software applications (http://www.cuahsi.org/his). Thus, ODM supports a set of functions that are not available through simple file-based data publishing.

[49] **Acknowledgments.** ODM has been developed as part of the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic Information System (HIS) project whose goal is to advance information system technology for hydrologic science (http://www.cuahsi.org/his.html). This work was supported by the National Science Foundation grants EAR 0413265 and EAR 0622374 for the development of Hydrologic Information Systems. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We gratefully acknowledge the comments and suggestions from those who have reviewed this data model. These comments have greatly enhanced its structure and function. In addition, we acknowledge the work of the CUAHSI HIS development team in assisting in the design and implementation of ODM.

References

Ahern, T. (2004), Earth Scope US Array data management plan, report, Inc. Res. Inst. for Seismol. Data Manage. Cent., Seattle, Wash. (Available at http://www.iris.edu/USArray/publications/US_Data_Plan_Final-V7.pdf)

Bandaragoda, C. J., D. G. Tarboton, and D. R. Maidment (2005), User needs assessment, in *Hydrologic Information System Status Report, Version 1*, edited by D. R. Maidment, chap. 4, pp. 48–87, Consortium of Univ. for the Adv. of Hydrol. Sci., Washington, D. C. (Available at http://www.cuahsi.org/docs/HISStatusSept15.pdf)

Bandaragoda, C., D. G. Tarboton, and D. R. Maidment (2006), Hydrology's effort towards the cyberfrontier, *Eos Trans. AGU*, *87*(1), 2–6, doi:10.1029/2006EO010005.

Blöschl, G., and M. Sivapalan (1995), Scale issues in hydrological modelling: A review, *Hydrol. Processes*, *9*, 251–290, doi:10.1002/hyp.3360090305.

Bose, R. (2002), A conceptual framework for composing and managing scientific data lineage, in *Proceedings of the 14th International Conference on Scientific and Statistical Database Management*, pp. 15–19, IEEE Press, Piscataway, N. J.

Bouganim, L., et al. (2001), The Ecobase Project: Database and Web technologies for environmental information systems, *SIGMOD Rec.*, *30*(3), 70–75, doi:10.1145/603867.603879.

Connolly, T., and C. Begg (2005), Database Systems: A Practical Approach to Design, Implementation, and Management, 4th ed., Addison-Wesley, Harlow, U. K.

Environmental Data Standards Council (2006), Environmental sampling, analysis, and results data standards: Overview of component data standards, *Stand. EX000001.1*, Environ. Data Stand. Counc., U. S. Environ. Prot. Agency, Washington, D.C. (Available at http://www.envdatastandards.net/files/693_file_ESAR_Overview_01_06_2006__Final_.pdf)

Goodall, J. L., J. S. Horsburgh, T. L. Whiteaker, D. R. Maidment, and I. Zaslavsky (2008), A first approach to Web services for the National Water Information System, *Environ. Model. Softw.*, *23*(4), 404–411, doi:10.1016/j.envsoft.2007.01.005.

Gray, J., D. T. Liu, M. Nieto-Santisteban, A. Szalay, D. J. DeWitt, and G. Heber (2005), Scientific data management in the coming decade, *SIGMOD Rec.*, *34*(4), 34–41, doi:10.1145/1107499.1107503.

Helsel, D. R. (1990), Less than obvious: Statistical treatment of data below the detection limit, *Environ. Sci. Technol.*, *24*(12), 1766–1774, doi:10.1021/es00082a001.

Maidment, D. R. (Ed.) (2002), *Arc Hydro GIS for Water Resources*, 203 pp., ESRI Press, Redlands, Calif.

Michener, W. K., J. W. Brunt, J. J. Helly, T. B. Kirchner, and S. G. Stafford (1997), Nongeospatial metadata for the ecological sciences, *Ecol. Appl.*, *7*(1), 330–342, doi:10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2.

National Water Quality Monitoring Council (2006), Water quality data elements: A user guide, *Tech. Rep. 3*, Advis. Comm. on Water Inf., Washington, D. C. (Available at http://acwi.gov/methods/pubs/wdqe_pubs/wqde_trno3.pdf)

Pokorný, J. (2006), Database architectures: Current trends and their relationships to environmental data management, *Environ. Model. Software*, *21*, 1579–1586, doi:10.1016/j.envsoft.2006.05.004.

Tomasic, A., and E. Simon (1997), Improving access to environmental data using context information, *SIGMOD Rec.*, *26*(1), 11–15, doi:10.1145/248603.248606.

J. S. Horsburgh and D. G. Tarboton, Utah Water Research Laboratory, Utah State University, 8200 Old Main Hill, Logan, UT 84322, USA. (jeff.horsburgh@usu.edu)

D. R. Maidment, Center for Research in Water Resources, University of Texas at Austin, Austin, TX 78712, USA.

I. Zaslavsky, San Diego Supercomputer Center, University of California, San Diego, MC 0505, 9500 Gilman Drive, La Jolla, CA 92093, USA.